SCIENTIFIC REPORTS

Correction: Author Correction

OPEN

Multi-label Deep Learning for Gene Function Annotation in Cancer Pathways

Renchu Guan^{1,3}, Xu Wang¹, Mary Qu Yang³, Yu Zhang^{1,4}, Fengfeng Zhou¹, Chen Yang⁵ & Yanchun Liang^{1,2}

Received: 31 August 2017

Accepted: 27 November 2017

Published online: 10 January 2018

The war on cancer is progressing globally but slowly as researchers around the world continue to seek and discover more innovative and effective ways of curing this catastrophic disease. Organizing biological information, representing it, and making it accessible, or biocuration, is an important aspect of biomedical research and discovery. However, because maintaining sophisticated biocuration is highly resource dependent, it continues to lag behind the continually being generated biomedical data. Another critical aspect of cancer research, pathway analysis, has proven to be an efficient method for gaining insight into the underlying biology associated with cancer. We propose a deep-learning-based model, Stacked Denoising Autoencoder Multi-Label Learning (SdaMLL), for facilitating gene multi-function discovery and pathway completion. SdaMLL can capture intermediate representations robust to partial corruption of the input pattern and generate low-dimensional codes superior to conditional dimension reduction tools. Experimental results indicate that SdaMLL outperforms existing classical multi-label algorithms. Moreover, we found some gene functions, such as Fused in Sarcoma (FUS, which may be part of transcriptional misregulation in cancer) and p27 (which we expect will become a member viral carcinogenesis), that can be used to complete the related pathways. We provide a visual tool (<https://www.keaml.cn/gpvisual>) to view the new gene functions in cancer pathways.

Cancer research has witnessed rapid advances year by year, generating a more abundant and complex body of knowledge. Researchers continue to come up with ingenious approaches for treating, preventing, and curing the disease. However, the war on cancer still has a long way to go¹. Among the 22 novel drugs approved by the U.S. Food and Drug Administration (FDA), six of them were designed for treating or diagnosing cancer². Tomas Lindahl and Paul Modrich's Mechanistic Studies of DNA Repair won the 2015 Nobel Prize in Chemistry, and their work may potentially advance the development of new cancer treatment³. In 2016, the 21st Century Cures Act provided 4.8 billion for the Cancer Moonshot⁴ and Precision Medicine Initiative, which aims at dramatically accelerating efforts to prevent, diagnose, and treat cancer⁵. Scientists from various fields, such as biology, statistics, and computer science, are using a vast array of approaches, trying their best to not only wage a battle but win the war against cancer worldwide.

Among these approaches, biocuration, which involves organizing, representing, and providing biological information for humans and computers, is an essential part of biomedical discovery and research⁶. At the present rate, however, the further and farther the curated data lags behind current biological knowledge, either way, the greater and more apparent the daily knowledge gaps will become. Originating with the Human Genome Project (HGP), microarray expression analysis, investments in large-scale sequencing centres and high-throughput analytical facilities have been increasing sharply, all leading to the exponential growth of biological data. The 2016 Nucleic Acids Research (NAR) online database collections, containing 15 categories and 41 subcategories, listed

¹Key Laboratory for Symbol Computation and Knowledge Engineering of National Education Ministry, College of Computer Science and Technology, Jilin University, Changchun, 130012, China. ²Zhuhai Laboratory of Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Zhuhai College of Jilin University, Zhuhai, 519041, China. ³MidSouth Bioinformatics Center and Joint Bioinformatics Ph.D. Program of University of Arkansas at Little Rock and Univ. of Arkansas Medical Sciences, Little Rock, AR, 72204, USA. ⁴Institute of Information Engineering, Chinese Academy of Sciences School of Cyber Security, University of Chinese Academy of Sciences, Beijing, 100093, China. ⁵College of Earth Sciences, Jilin University, Changchun, 130061, China. Renchu Guan and Xu Wang contributed equally to this work. Correspondence and requests for materials should be addressed to C.Y. (email: yangc616@jlu.edu.cn) or Y.L. (email: yliang@jlu.edu.cn)

Figure 1. Feature matrix generation flowchart. The upper half illustrates the process for extracting all the gene names from the pathways in KEGG¹² and the other half shows how we select articles that embed descriptions about gene function.

1664 published biological databases. By July 2017, more than 27 million citations for biomedical literature from MEDLINE, life science journals, and online books had been indexed in Pubmed⁷. However, the resources for generating and testing hypotheses will soon become depleted or ineffective because of gap expansions. The purpose of providing a greater amount of instantaneous manual annotation associated with increased data acquisition, while being prepared to address the possibility of having to make best use of purely human labour, creates a virtually insurmountable dilemma⁸. It is because this approach is totally dependent on well-trained professional biocurators who can analyse and extract categorized information from the published literature.

Text mining denotes the process of deriving high-quality information from text through the devising of patterns and trends. When properly processed by text mining methods, biomedical data may provide invaluable information to affect peoples' cognizance of biomedical phenomena⁹. In 2005, most text mining tools were suited merely to a limited number of tasks¹⁰. With more yearly biomedical challenges and newly published databases such as i2b2, TREC Medical/CDS and BioNLP, biomedical text mining techniques have been driven to progress sharply. Hisschman *et al.* list the following four requirements for a biocurator text mining tool¹¹: 1. Easy to use, install and maintain by the intended end user; 2. The tool need not be perfect, but it needs to complement the biocurator's function; 3. Initial batch processing is necessary; and 4. The tool should provide linking of gene expression mentions of biological entities identified in the text with their referents identified in biological databases, then link them to the appropriate ontological terms.

Pathway analysis is an efficient method for gaining insight into the underlying biology of differentially expressed genes and proteins with less complexity and better explanatory power. KEGG PATHWAY¹² collects manually drawn pathway maps integrating information from metabolism, genetic information processing, environmental information processing, cellular processes, organismal systems and human diseases. The workflow of biological pathway construction consists of four steps¹³ that serve as mining information resources, using pathway building tools, refinement, and leading to the desired specific annotated pathway, a process that can be very time-consuming and costly. Therefore, computational methods are needed to characterize the pathways automatically. Traditional *in silico* pathway annotation methods still rely on biologists to properly define features and guide feature selection during pathway prediction¹⁴, which creates yet another time- and cost-consuming step. Moreover, the results are highly dependent on the selected features, and if the features are not good enough, the achieved result may not be satisfying.

Intended for learning multiple levels of feature composition, models based on deep learning have become state-of-the-art methods in thriving research areas such as image recognition^{15–17}, speech recognition^{18–20}, language translation^{6,21,22}, sentiment analysis^{23–25}, image caption^{26–28} and so on. Deep learning is a powerful tool for discovering intricate structures in high-dimensional data²⁹, which is common in biomedical informatics.

In this paper, faced with the task of sifting through the voluminous extant biomedical publications, we propose a Stacked Denoising Autoencoder Multi-Label Learning (SdaMLL) model to explore the effects, if any, gene multi-functions may have on cancer pathways in KEGG. To acquire more functions for each gene, we explore full texts of biomedical articles where more detailed methodologies, experimental results, critical discussions and interpretations can be found³⁰. To the best of our knowledge, this is the first work applying a deep learning model³¹ to analyse gene multi-functions relevant to cancer pathways derived from full-text biomedical publications. In addition, the entire procedure proposed in this study does not require the involvement of a biologist to do a feature study about the data. Experimental results on eight KEGG cancer pathways reveal that SdaMLL is not only superior to classical multi-label learning models such as K-nearest neighbours and decision tree, but it can also achieve numerous gene functions related to important cancer pathways.

Material and Methods

Faced with the challenge of reconciling a tremendous body of biomedical literature with its biological referents, we deem deep learning to be one of the most promising methods for fleshing out the relationship between them, and particularly in this paper, relationships with genes affecting cancer pathways. To predict multi-functions for each gene, we first generate a feature matrix following the routine in Fig. 1.

Figure 2. The architecture of SdaMML. Each row of the feature matrix represents a gene. After being fed into the stacked denoising autoencoder, the original vector is tuned by removing various noises. The output of the autoencoder is then provided as the input to BP-MLL, predicting the gene assignment to the pathways.

To generate the feature matrix, as shown in Fig. 1, on the one hand, we first extract all the gene IDs from a given KGML file. A KGML file provides information on reaction objects and their interactions annotated in the KEGG pathway plots, and the orthologous gene annotations from the KEGG GENES database. Widespread biomedical ambiguity is a well-known issue in bioinformatics. Therefore, we try to identify and tag all of the salient referents of a particular gene. Through the KEGG Entry list API, the most frequently used names for each gene in all the eight pathways are successfully gathered. On the other hand, with the help of Gene_pubmed_rif relation assembled in the ELink utility provided by NCBI, the PubMed id of all articles covering specific gene functions are fetched. Since a small portion of the articles are not open access, articles possessing both PubMed id and PMC id in a text file containing the list of all the downloadable pdfs via PMC FTP Service are left remaining. After downloading all 18930 pdf files, we extracted the required text content. The resulting matrix represents genes with term frequencies.

After generating the feature matrix, the deep learning-based method SdaMML can unveil gene multi-function. As shown in Fig. 2, SdaMML consists of two modules: (a) Stacked denoising autoencoders (SdAs) serve as the representation learner, capturing the dependencies between dimensions in high dimensional distribution, and (b) the backpropagation for Multi-Label Learning (BP-MLL) focuses on finding the proper pathway label for each gene.

A traditional autoencoder involves encoding and decoding. Encoding maps the input $x \in [0, 1]^d$ to a hidden representation $y \in [0, 1]^{d'}$ through a deterministic mapping:

$$y = s_{\theta}(\mathbf{W}x + \mathbf{b}) \quad (1)$$

where $\theta = \{\mathbf{W}, \mathbf{b}\}$ and s denotes the sigmoid function. The code y is a latent representation of x , and it is then mapped back into z , which has the same shape as x . The reconstruction mapping process is formulated using a similar transformation:

$$z = s_{\theta'}(\mathbf{W}'y + \mathbf{b}') \quad (2)$$

where $\theta' = \{\mathbf{W}', \mathbf{b}'\}$. The autoencoder is constrained to have tied weights, which means $\mathbf{W} = \mathbf{W}'$. The goal of auto-encoder is to minimize the average reconstruction error:

$$\begin{aligned} \theta^*, \theta'^* &= \arg \min_{\theta, \theta'} \frac{1}{n} \sum_{i=1}^n L(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}) \\ &= \arg \min_{\theta, \theta'} \frac{1}{n} \sum_{i=1}^n L(\mathbf{x}^{(i)}, s_{\theta'}(s_{\theta}(\mathbf{x}^{(i)}))) \end{aligned} \quad (3)$$

For purposes of extending the hidden layer's ability to discover more robust features and avoiding learning the identity simply, the denoising autoencoder is trained from a corrupted version of input. The denoising autoencoder is a stochastic version of an autoencoder, i.e., the initial input $\mathbf{x}$ is stochastically mapped to $\tilde{\mathbf{x}}$ with

$\tilde{\mathbf{x}} \sim q_D(\tilde{\mathbf{x}}|\mathbf{x})$. As shown in Fig. 2, for each input $\mathbf{x}$, a fixed number vd of elements is randomly chosen to be reset as 0, leaving others unchanged. The corrupted version of the original input, $\tilde{\mathbf{x}}$, is then mapped to a hidden representation from which we reconstruct. The mapping and reconstruction processes are the same as those performed in a typical autoencoder. The goal of the denoising autoencoder is to minimize the reconstruction error $L_H(\mathbf{x}, \mathbf{z}) = \mathbb{H}(\mathcal{B}_x|\mathcal{B}_z)$ over the training set as well. A multi-layer denoising autoencoder could be constructed by combining multiple single-layer denoising autoencoders and connecting the output of the previous layer to the input of the next one. Each layer may be tuned by an unsupervised pre-training. Once the first k layers are trained, we can train the $(k+1)^{th}$ layer because we can now compute the code or latent feature representation from the layer below.

The upper right part of Fig. 2 illustrates the framework of BP-MLL, which is a typical feed-forward artificial neural network modified for multi-label learning. BP-MLL solves the problem with two notable modifications: (a) a specifically designed error function and (b) a revision made according to the classical learning algorithm. It outperforms several existing methods in functional genomics and text categorization³². Let $\chi = \mathbb{R}^d$ denote the instance domain and $\gamma = 1, 2, \dots, N$, indicates the set of all the class labels. m indicates multi-label instances, and each Y_i in $\{(x_1, Y_1), (x_2, Y_2), \dots, (x_m, Y_m)\}$ may contain several labels. The global error function is rewritten as

$$E = \sum_{i=1}^m E_i = \sum_{i=1}^m \frac{1}{|Y_i| |\bar{Y}_i|} \sum_{(k,l) \in Y_i \times \bar{Y}_i} \exp(-(q_k^i - q_l^i)) \quad (4)$$

where for the i -th training example (x_i, Y_i) , the error term is $\frac{1}{|Y_i| |\bar{Y}_i|} \sum_{(k,l) \in Y_i \times \bar{Y}_i} \exp(-(q_k^i - q_l^i))$. The complementary set of Y_i in γ is $\bar{Y}_i$ and $|\cdot|$ is the cardinality of a set. The difference between the outputs of BP-MLL on one label belonging to $x_i (k \in Y_i)$ and one label not belonging to it ($l \in \bar{Y}_i$) is measured by $c_k^i - c_l^i$. The performance gets better when the difference increases. Gradient descent strategy is applied to reduce the error:

$$\begin{aligned} \Delta_{w_{sj}} &= -\alpha \frac{\partial E_i}{\partial w_{sj}} = -\alpha \frac{\partial E_i}{\partial netc_j} \frac{\partial netc_j}{\partial w_{sj}} \\ &= \alpha d_j \left[\frac{\delta \left(\sum_{s=1}^M b_s w_{sj} \right)}{\partial w_{sj}} \right] = \alpha d_j b_s \end{aligned} \quad (5)$$

$$\begin{aligned} \Delta_{v_{hs}} &= -\alpha \frac{\partial E_i}{\partial v_{hs}} = -\alpha \frac{\partial E_i}{\partial netb_s} \frac{\partial netb_s}{\partial v_{hs}} \\ &= \alpha e_s \left[\frac{\delta \left(\sum_{h=1}^d a_h v_{hs} \right)}{\partial v_{hs}} \right] = \alpha e_s a_h \end{aligned} \quad (6)$$

the bias are changed according to:

$$\Delta \theta_j = \alpha d_j; \quad \Delta \gamma_s = \alpha e \quad (7)$$

where α is the learning rate and its range is (0.0, 1.0). When training the BP-MLL, the training instances are fed to the network one by one. For each multi-labeled instances (x_i, Y_i) , the weights are updated according to Eqs 5 and Eqs 7. When global error E doesn't decrease or the training epochs hits a threshold, it stops training.

Results and Discussion

The data used in our experiment is downloaded from the two databases KEGG PATHWAY and PubMed Central. Term frequencies are computed for all the words appearing in pathway-related articles retrieved from PubMed Central. To construct the feature matrix, we collect the gene names for all 1144 genes and 18930 articles closely related to these genes. The feature matrix is constructed using all the genes listed in the pathway map overviews and gene-related full-text articles fetched from Pubmed. For purposes of validating the generalization ability of the proposed method, 10-fold cross-validation was conducted while only being able to use a limited number of genes extracted from the existing KEGG pathways. Regarding the pathways, we selected the cancer pathways according to the following widely influential references. Douglas *et al.* summarized six capabilities acquired by most forms of cancer, and they added two more cancer traits in 2011^{33,34}. Parallel pathways take part in the tumorigenesis process³⁵⁻³⁷. During the process, traits of cancers can be arranged in multiple permutations, which means cell transformation occurs. Semir Beyaz *et al.* have proposed that uncovering how PPAR- δ mediates tumorigenesis in diverse tissues and cell types in response to diet may enhance clinical utility³⁸. Niall *et al.* regard targeting DNA repair and repair-deficient tumours as new avenues for treating advanced disease in the future³⁹.

We compare SdaMLL with the following methods: K-nearest neighbours (KNN), decision trees (DTs)⁴⁰ and Backpropagation for Multi-Label Learning (BP-MLL)³². KNN is one of the most commonly used machine learning methods. The goal of the nearest neighbour method is to find a pre-determined number of training samples closest in distance to the new point and predict the labels among these samples. DTs is a non-parametric supervised learning classification method that aims at learning simple decision rules derived from original data and predicting results with these rules. BP-MLL is derived from the backpropagation⁴¹ algorithm via applying an error function acquiring the characteristics of multi-label learning. This is the first time this term has been employed in functional genomics and text categorization.

	Coverage precision	Ranking loss	Coverage
KNN	0.333	0.662	95.25
Decision trees	0.385	0.644	100.763
SdaMML	0.577	0.286	2.436
BP-MML	0.529	0.306	2.608

Table 1. Experiments results for all multi-label classification algorithm. Results of BP-MML and SdaMML with highest average precision are selected.

In our experiment, we adopted three different metrics for multi-label learning proposed by Schapire et. al, which are coverage, ranking loss, and average precision⁴². Coverage evaluates the performance of a system for the top-ranked label, i.e., how dire is the need to go further down the set of labels to cover all the correct labels of each instance, on average. The performance is improved when the value decreases. For a set of labeled documents $S = \langle (x_1, Y_1), \dots, (x_m, Y_m) \rangle$, the coverage is

$$coverage_s(H) = \frac{1}{m} \sum_{i=1}^m \max_{\ell \in Y_i} rank_f(x_i, \ell) - 1 \quad (8)$$

Ranking loss evaluates the fraction of label pairs in which an irrelevant label is ranked higher than a relevant label. When the $rloss(H) = 0$, the performance is perfect. For a set of labeled documents $S = \langle (x_1, Y_1), \dots, (x_m, Y_m) \rangle$, ranking loss is defined as

$$rloss(H) = \frac{1}{m} \sum_{i=1}^m \frac{|D_i|}{|Y_i| |\bar{Y}_i|} \quad (9)$$

where $\bar{Y}$ is the complementary set of Y in γ and $D_i = \{(y_1, y_2) | f(x_i, y_1) \leq f(x_i, y_2), (y_1, y_2) \in Y_i \times \bar{Y}_i\}$

The average precision evaluates the average fraction of labels ranked above a particular label $y \in Y$, which actually are in Y . The performance is better when the value of average precision is bigger.

$$avg\ pres_s(H) = \frac{1}{m} \sum_{i=1}^m \frac{1}{|Y_i|} \sum_{\ell \in Y_i} \frac{|\{\ell' \in Y_i | rank_f(x_i, \ell') \leq rank_f(x_i, \ell)\}|}{rank_f(x_i, \ell)} \quad (10)$$

1114 distinct genes are distributed in 8 pathways, which leads to a relatively small amount of data. Under such circumstances, we try to check the generalization ability of SdaMML and other methods with 10-fold cross-validation. During the experiments, the dataset is randomly divided into 10 separate subsets with only one subset serving as validation in each assessment.

Results for average precision can be found in the second column in Table 1, and Fig. 3(a) elucidates a detailed comparison between SdaMML and BP-MML. SdaMML achieves the best performance among the 4 algorithms. SdaMML's average precision is 0.577, while KNN's average precision is 0.333. Compared to KNN, SdaMML's is 73.27% higher. BP-MML's average precision is the second highest when confronted with the other algorithms. Nevertheless, to reach such a high average precision, BP-MML requires 40 more training epochs than SdaMML. In contrast, SdaMML converges with fewer than 10 training epochs.

In the third column (Ranking Loss) of Table 1, we can see the value of ranking loss, and a detailed comparison between BP-MML and SdaMML appears in Fig. 3(b). SdaMML's ranking loss is 0.286, which is 0.376 lower than KNN, and which represents the best performance among the 4 algorithms. The results are uniformly comparable to the values in the second column; KNN and decision trees get similar results; and SdaMML and BP-MML's ranking loss values are alike. When BP-MML is compared with SdaMML, the difference becomes increasingly clear. SdaMML, again, converges much faster than BP-MML.

The forth column (Coverage) of Table 1 has the coverage values. Figure 3(c) shows a comparison between BP-MML and SdaMML. SdaMML achieves the best performance and remains stable during the training no matter how variable the number of epochs is. The coverage of SdaMML, 2.436, is relatively low, constituting a significant signal that SdaMML will not go deeply into the set of labels when determining correct labels for particular instances. In the meantime, while the result of BP-MML's coverage is approximately low, it is 0.172 higher than SdaMML. However, the performance disparity between SdaMML and decision trees is very spectacular. Specifically, decision tree's convergence is virtually 40 times greater than that of SdaMML.

For the time consumption, it costs 153.07 s to train the BP-MML model per epoch while the time of training the SdaMML model is averaged to 38.46 s per epoch. The training time for KNN and decision tree is 0.42 s and 6.63 s. All the prediction time of KNN and decision tree is 12.6 s and 16ms, however, BP-MML and SdaMML is 7.16 s and 4.56 s. From the prediction time comparison, it can be seen that SdaMML needs more time than decision trees, however, it is faster than KNN.

Based on the results of the aforementioned measurements, SdaMML outperforms BP-MML in three respects: (1) SdaMML converges far more quickly than BP-MML, requiring 30 fewer training epochs on average; (2) On all of the 3 metrics, SdaMML achieves better results than BP-MML; and, (3) Original BP-MML is time-consuming, while SdaMML finishes in less time. Although the process of constructing the feature matrix is complicated, when one considers the difference between the training data distribution and the actual data distribution, the training data distribution is probably corrupted. Additionally, each instance is a 1×18930 vector, and most of the bits in

Figure 3. The Comparison between BP-MLL and SdaMLL. $\uparrow$ denotes that the model's performance is better when the metrics are larger and vice versa.

a particular vector are zero, creating a data sparsity problem in which the data are extremely hard to interpret. Stacked denoising autoencoders, on the one hand, try to capture intermediate representation, which is robust compared to a partial corruption of the input pattern. In addition, by initializing the weights effectively, the architecture of the autoencoder is able to generate low-dimensional codes that surpass common dimension reduction tools such as principal component analysis⁴³ and independent component analysis⁴⁴ (ICA).

Apart from the quantity improvements, we can also ascertain the biological results. As mentioned in the dataset description, all the collected articles indicate particular functions of genes, which can serve as evidence for judging whether the gene serves as part of a pathway. From this prediction, we also found some articles that can solidly back up the correlation between genes and pathways. First, tumour Protein P53 (TP53) encodes a tumour suppressor protein containing transcriptional activation, DNA binding, and oligomerization domains. The encoded protein responds to diverse cellular stresses to regulate expression of target genes, thereby inducing cell cycle arrest, apoptosis, senescence, DNA repair, or changes in metabolism. Saha *et al.* demonstrated a possible role for the candidate tumour suppressor ING genes in the biology of EBV-associated cancer⁴⁵, and the N-terminal domain of EBNA3C residues 129 to 200 was previously demonstrated to associate with p53. From these clues, we may regard p53 as part of the pathway describing pathways in cancer (hsa:05200).

Moreover, FUS or FUS/TLS (Fused in sarcoma/translocated in liposarcoma) is a multifunctional RNA/DNA-binding protein that is pathologically associated with cancer and neurodegeneration. It encodes a multifunctional protein component of the heterogeneous nuclear ribonucleoprotein (hnRNP) complex. From an article⁴⁶, we can conclude that FUS-DDIT3 deregulates some NF-kappaB-controlled genes through interaction with NFKBIZ. FUS related genes are involved in tumour type-specific fusion oncogenes in human malignancies, which indicates that FUS is probably part of transcriptional misregulation in cancer (hsa:05202).

In addition, P27 or Kip1 (Cyclin-Dependent Kinase Inhibitor 1B) encodes a cyclin-dependent kinase inhibitor, in which mutations are associated with multiple endocrine neoplasia type IV. The encoded protein binds to and prevents the activation of cyclin E-CDK2 or cyclin D-CDK4 complexes, and thus controls cell cycle progression at G1. PKB/Akt mediates cell-cycle progression by phosphorylation of P27 (Kip1) at threonine 157, and modulation of its cellular localization indicates that Akt may contribute to tumour-cell proliferation by phosphorylation and cytosolic retention of p27, thus relieving CDK2 from p27-induced inhibition⁴⁷. Obviously, it is also a hint that p27 becomes a member of viral carcinogenesis (hsa:05203). Other findings such as CCNA2, EGR2, and MAPKAPK2 belong to viruses or encode cancer pathways; CCND1, NRAS, and SDC1 all contribute to the proteoglycans pathway; AKT1, MET, TP53 and MAPK1 are adaptations of cellular metabolism pathway elements.

These three samples are the solid proof for our biology discovery. More interesting examples (270 new functions) are listed at <https://www.keaml.cn/gpvisual>.

Figure 4. Illustration of new gene functions. **(a)** Network connecting genes, articles and pathways. The yellow points are the pathways, light blue points are the PubMed manuscripts and blue points are functional genes. **(b)** Detailed description about relations. All the relationships are listed as triple-items {gene, pathway, article}. **(c)** Predicted results. The illustration of our predicted results about gene functions. **(d)** Visualizing predicted genes in KEGG. The completion results for KEGG pathways¹² are listed at the bottom of the web page.

We also provide a visualization tool for multi-function genes in cancer pathways. Based on the text mining results, we find it is common for a gene to get involved in multiple pathways. For the sake of unveiling the relationship between genes and pathways, we try to visualize a network of genes, articles, and pathways. In Fig. 4(a), we can observe a primitive network illustrating help edges connecting genes and pathways. If users click on the info button, the helper information containing a definition for each entity and lines in the network will either appear or hide. In Fig. 4(b), specifics about each link can be found in the details tab. Users can input the desired name or ID of specific genes to dig out relations between those genes and pathways, as depicted in Fig. 4(c). Meanwhile, thumbnails of predicted pathways will pop out. If users click on the thumbnails, the browser will jump to a detailed description of that pathway, as shown in Fig. 4(d). In the new tab, predicted results are listed at the bottom of the web page.

Conclusion and Future Work

In this paper, we have proposed a novel multi-label gene function annotation model based on a deep learning strategy, namely, SdaMLL, for gene multi-function discovery. This model takes advantage of both effective dimension reduction and multi-label classification on account of Stacked denoising autoencoders. Compared to BP-MLL, SdaMLL converges much faster in terms of the number of training epochs. In addition, during the experiments, we try to reduce the dimension from 18900 to 200, which helps to shorten the training time tremendously. From the results, we can conclude that SdaMLL is a state-of-the-art algorithm for finishing the task at hand. In addition, we provide a website for researchers to inspect relationships between genes and articles.

This study demonstrated how the proposed method performed based on the data of eight pathways and generated a feature matrix containing genes existing in these pathways. Better annotation performance may be anticipated if more information from other pathways is integrated into the model in the future.

References

1. Hanahan, D. Rethinking the war on cancer. *The Lancet* **383**, 558–563 (2014).
2. Jenkins, J. A Review of CDER's Novel Drug Approvals for 2016. <https://blogs.fda.gov/fdavoices/index.php/2017/01/a-review-of-cders-novel-drug-approvals-for-2016/> (2017).
3. Nobelprize.org. The Nobel Prize in Chemistry 2015. https://www.nobelprize.org/nobel_prizes/chemistry/laureates/2015/ (2017).
4. Biden, J. Inspiring a New Generation to Defy the Bounds of Innovation: A Moonshot to Cure Cancer. <https://medium.com/cancer-moonshot/inspiring-a-new-generation-to-defy-the-bounds-of-innovation-a-moonshot-to-cure-cancer-fbd71d01c2e#dq2us5l9w> (2016).

5. DeBonis, M. Congress passes 21st Century Cures Act, boosting research and easing drug approvals. <https://www.washingtonpost.com/news/powerpost/wp/2016/12/07/congress-passes-21st-century-cures-act-boosting-research-and-easing-drug-approvals/> (2016).
6. Howe, D. *et al.* Big data: The future of biocuration. *Nature* **455**, 47–50 (2008).
7. PubMed. Home - PubMed - NCBI. <https://www.ncbi.nlm.nih.gov/pubmed/> (2017).
8. Burge, S. *et al.* Biocurators and biocuration: surveying the 21st century challenges. *Database* **2012**, bar059 (2012).
9. Zhai, X. *et al.* Research status and trend analysis of global biomedical text mining studies in recent 10 years. *Scientometrics* **105**, 509–523 (2015).
10. Rebholz-Schuhmann, D., Kirsch, H. & Couto, F. Facts from text—is text mining ready to deliver? *PLoS Biology* **3**, e65 (2005).
11. Hirschman, L. *et al.* Text mining for the biocuration workflow. *Database* **2012**, bas020 (2012).
12. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. Kegg as a reference resource for gene and protein annotation. *Nucleic Acids Research* **44**, D457–D462 (2016).
13. Viswanathan, G. A., Seto, J., Patil, S., Nudelman, G. & Sealfon, S. C. Getting started in biological pathway construction and analysis. *PLoS Computational Biology* **4**, e16 (2008).
14. Dale, J. M., Popescu, L. & Karp, P. D. Machine learning methods for metabolic pathway prediction. *BMC Bioinformatics* **11**, 15 (2010).
15. Ciregan, D., Meier, U. & Schmidhuber, J. Multi-column deep neural networks for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 3642–3649 (IEEE, 2012).
16. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).
17. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
18. Graves, A., Mohamed, A.-r. & Hinton, G. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 IEEE International Conference on*, 6645–6649 (IEEE, 2013).
19. Hinton, G. *et al.* Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine* **29**, 82–97 (2012).
20. Dahl, G. E., Yu, D., Deng, L. & Acero, A. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions Audio, Speech, Language Processing* **20**, 30–42 (2012).
21. Sutskever, I., Vinyals, O. & Le, Q. V. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 3104–3112 (2014).
22. Johnson, M. *et al.* Google's multilingual neural machine translation system: enabling zero-shot translation. *arXiv preprint arXiv:1611.04558* (2016).
23. Zhou, S., Chen, Q. & Wang, X. Active deep learning method for semi-supervised sentiment classification. *Neurocomputing* **120**, 536–546 (2013).
24. Tang, D., Wei, F., Qin, B., Liu, T. & Zhou, M. Coooolll: A deep learning system for Twitter sentiment classification. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 208–212 (2014).
25. Socher, R. *et al.* Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, 1631–1642 (2013).
26. Vinyals, O., Toshev, A., Bengio, S. & Erhan, D. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3156–3164 (2015).
27. Chen, X. & Zitnick, C. L. Learning a recurrent visual representation for image caption generation. *arXiv preprint arXiv:1409.2329* (2014).
28. Zaremba, W., Sutskever, I. & Vinyals, O. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329* (2014).
29. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
30. Guan, R., Yang, C., Marchese, M., Liang, Y. & Shi, X. Full text clustering and relationship network analysis of biomedical publications. *PloS one* **9**, e108847 (2014).
31. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y. & Manzagol, P.-A. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research* **11**, 3371–3408 (2010).
32. Zhang, M.-L. & Zhou, Z.-H. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Transactions Knowledge and Data Engineering* **18**, 1338–1351 (2006).
33. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).
34. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
35. Cully, M., You, H., Levine, A. J. & Mak, T. W. Beyond pten mutations: the pi3k pathway as an integrator of multiple inputs during tumorigenesis. *Nat. Reviews Cancer* **6**, 184–192 (2006).
36. D'Cruz, C. M. *et al.* c-myc induces mammary tumorigenesis by means of a preferred pathway involving spontaneous kras2 mutations. *Nat. Medicine* **7**, 235–239 (2001).
37. Kolli, F. T., Bommer, G. & Göke, B. Wnt/beta-catenin/tcf signaling: a critical pathway in gastrointestinal tumorigenesis. *Digestion* **66**, 131–144 (2002).
38. Beyaz, S. & Yilmaz, Ö. H. Molecular pathways: dietary regulation of stemness and tumor initiation by the ppar- δ pathway. *Clinical Cancer Research* **22**, 5636–5641 (2016).
39. Corcoran, N. M., Clarkson, M. J., Stuchbery, R. & Hovens, C. M. Molecular pathways: Targeting dna repair pathway defects enriched in metastasis. *Clinical Cancer Research* **22**, 3132–3137 (2016).
40. Quinlan, J. R. Induction of decision trees. *Machine Learning* **1**, 81–106 (1986).
41. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. *et al.* Learning representations by back-propagating errors. *Cognitive modeling* **5**, 1 (1988).
42. Schapire, R. E. & Singer, Y. Boostexter: A boosting-based system for text categorization. *Machine Learning* **39**, 135–168 (2000).
43. Wold, S., Esbensen, K. & Geladi, P. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems* **2**, 37–52 (1987).
44. Comon, P. Independent component analysis, a new concept? *Signal Processing* **36**, 287–314 (1994).
45. Saha, A., Bamidele, A., Murakami, M. & Robertson, E. S. Ebn3c attenuates the function of p53 through interaction with inhibitor of growth family proteins 4 and 5. *Journal of Virology* **85**, 2079–2088 (2011).
46. Göransson, M. *et al.* The myxoid liposarcoma fus-ddit3 fusion oncoprotein deregulates nf- κ b target genes by interaction with nfkbiz. *Oncogene* **28**, 270–278 (2009).
47. Shin, I. *et al.* Pkb/akt mediates cell-cycle progression by phosphorylation of p27kip1 at threonine 157 and modulation of its cellular localization. *Nat. Medicine* **8**, 1145–1152 (2002).

Acknowledgements

The authors are grateful for the support of the National Natural Science Foundation of China (No.61572228, No.61472158, No.61300147, No.61602207), United States National Institutes of Health (NIH) Academic Research Enhancement Award (No.1R15GM114739), National Institute of General Medical Sciences (NIH/NIGMS) (No.5P20GM103429), United States Food and Drug Administration (FDA) (No.HHSF223201510172C), the

Science Technology Development Project from Jilin Province (No.20160101247JC), Zhuhai Premier-Discipline Enhancement Scheme and Guangdong Premier Key-Discipline Enhancement Scheme. This work was partially supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB13040400) and a start-up grant from the Jilin University. We thank KEGG group for supporting pathway data.

Author Contributions

This study was conceived by R.C.G., C.Y. and Y.C.L. X.W. and Y.Z. performed experiments. R.C.G., X.W., M.Q.Y. and F.F.Z. analyzed the data. R.C.G. and X.W. drafted the manuscript. All the authors revised the manuscript and approved the final manuscript.

Additional Information

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2017